BMC Bioinformatics

BioMed Central

Oral presentation

Open Access

Discovering relationships among dispersed repeats using spatial association rule mining

Surya Saha*[1,2,3], Susan Bridges[1,2], Zenaida Magbanua[1,3,4] and Daniel G Peterson[1,3,4]

Address: [1]Institute for Digital Biology, Mississippi State University, Mississippi State, MS 39762, USA, [2]Department of Computer Science and Engineering, Mississippi State University, Mississippi State, MS 39762, USA, [3]Mississippi Genome Exploration Laboratory, Mississippi State University, Mississippi State, MS 39762, USA and [4]Department of Plant and Soil Sciences, Mississippi State University, Mississippi State, MS 39762, USA

Email: Surya Saha* - ss307@cse.msstate.edu

* Corresponding author

from Fourth International Society for Computational Biology (ISCB) Student Council Symposium
Toronto, Canada. 18 July 2008

Published: 30 October 2008

BMC Bioinformatics 2008, **9**(Suppl 10):O4 doi:10.1186/1471-2105-9-S10-O4

This abstract is available from: http://www.biomedcentral.com/1471-2105/9/S10/O4

Background

DNA in eukaryotic genomes is characterized, and often dominated, by repetitive, non-genic DNA sequences. Initially thought to be non-functional, repeats have been found to influence gene expression [1] and provide diversity to the genome via mutation. Mobile repeat sequences [2](transposons) have played a prominent role in the evolutionary histories of eukaryotic genomes [3,4], and their persistence in eukaryotic DNA indicates that they have, on the whole, been evolutionarily advantageous. While there are an increasing number of algorithms that have been developed for discovering novel dispersed repeats [5-7], significant analysis of the repeats and their relationships to other genome features will be required before we can truly understand the complex ways in which dispersed repeat sequences contribute to evolutionary fitness. We propose a spatial proximity rule based data mining technique to discover highly fragmented repeat regions for which only the conserved parts are reported by a computational repeat finder.

Materials and methods

We present an algorithm for mining the coordinates of different families of *ab initio* identified repetitive regions on chromosomal length DNA sequences to yield proximity relationships between repeat families [8]. Association rule mining [9] is used to compute the statistical significance of the discovered relationships. False positives are screened out by means of Monte Carlo methods. The filtered proximity relationships are in turn used to build graphs in which repeat families correspond to the vertices and the discovered proximity relationships correspond to edges. Connected components are extracted from the graphs to yield sets of related families denoting diverged repeat regions.

Results and conclusion

We demonstrate that this approach applied to the rice genome [10] can discover annotated repeat regions [11,12] and can identify novel relationships among repetitive DNA sequences. The novel relationships can be used to detect hitherto unknown repeat regions in sequenced genomes. The approach described can be extended to address and investigate proximity relationships between all annotated elements within a genome including genes, repetitive elements, non-coding RNAs, and regulatory elements.

References

1. Dorer DR, Henikoff S: **Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila.** *Cell* 1994, **77**:993-1002.
2. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, *et al.*: **A unified classifica-**

tion system for eukaryotic transposable elements. *Nat Rev Genet* 2007, **8(12):**973-982.
3. McClintock B: **The significance of responses of the genome to challenge.** *Science* 1984, **226(4676):**792-801.
4. Biemont C, Vieira C: **Genetics: Junk DNA as an evolutionary force.** *Nature* 2006, **443(7111):**521-524.
5. Bergman CM, Quesneville H: **Discovering and detecting transposable elements in genome sequences.** *Briefings in Bioinformatics* 2007, **8(6):**bbm048.
6. Saha S, Bridges S, Magbanua Z, Peterson D: **Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences.** *Tropical Plant Biology* 2008, **1(1):**85-96.
7. Saha S, Bridges S, Magbanua ZV, Peterson DG: **Empirical comparison of ab initio repeat finding programs.** *Nucleic Acids Research* 2008, **36(7):**2284-2294.
8. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21(1):**i351-i358.
9. Agrawal R, Imielinski T, Swami A: **Mining association rules between sets of items in large databases.** In *ACM SIGMOD Conference Washington D.C.: ACM Press New York, NY, USA*; 1993:207-216.
10. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, *et al.*: **The TIGR Rice Genome Annotation Resource: improvements and new features.** *Nucleic Acids Res* 2007, **35(Database issue):**D883-D887.
11. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
12. Ouyang S, Buell CR: **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants.** *Nucleic Acids Res* 2004, **32(Database issue):**D360-D363.